Instituto de Estadística
Ingeniería en Estadística y Ciencia de Datos

Comparación de distintos modelos lineales generalizados de la familia Gamma para un problema con variable de respuesta sobre precios de casas en Ames, EE.UU.

Ricardo Menares
ricardomenares1@gmail.com

Valparaíso, Chile
5 de diciembre de 2022

Índice general

Capítulo 1

Introducción

La venta y compra de casas es una actividad que a menudo puede ser problemática, pues definir el precio de estas se basa en muchos factores, que en ocasiones algunos son hasta desconocidos por las personas. Si es cuestionable lo anterior, predecir el precio de casas lo será aún más, pues según Muralidharan & et al (2018) pronosticar precios de casas es difícil, ya que los factores que afectan el mercado de la vivienda van desde socioeconómicos (tasa de criminalidad per cápita, acceso al transporte, promedio ingreso y nivel educativo, entre otros) y características específicas de la casa (por ejemplo, metros cuadrados, número de dormitorios, estilo de casa, fecha de última remodelación), esto indica una necesidad de construir y aplicar distintos modelos a modo de que este desarrollo y la disponibilidad de varios modelos de predicción del precio de viviendas pueden influir un papel útil para llenar un vacío de información que puede mejorar la eficiencia del mercado inmobiliario.

Es por lo anterior, que en esta investigación se busca poder ajustar y comparar distintos modelos lineales generalizados de la familia Gamma, para la predicción del precio de viviendas. Para lograr lo anterior, se usa una base de datos que contiene varias variables de características de casas de la ciudad de Ames, Estados Unidos, incluida una variable del precio de estas. Luego de realizar un análisis exploratorio de datos, se aplican los distintos modelos lineales generalizados Gamma que se basan en 3 diferentes modelos con distinto enlace: Canónico, inverso y logarítmico.

1.1. Objetivo general

Ajustar y comparar un modelo lineal generalizado Gamma, con distintas funciones de enlace, para el ajuste del precio de casas en Ames, EE.UU. a través de distintas variables que caracterizan una casa y que puedan afectar al valor de esta.

1.2. Objetivos específicos

- Examinar el modelo teórico elegido para los datos obtenidos.
- Realizar un análisis exploratorio sobre los datos adquiridos.
- Ajustar el modelo lineal generalizado con respuesta gamma con distintas funciones de enlace para los datos propuestos.
- Comparar experimentalmente los diversos prototipos planteados, mediante análisis residual y de bondad de ajuste.
- Seleccionar el mejor modelo y modelar el precio de las casas.

Capítulo 2

Estado del arte

Existe una diversidad de estudios pasados que han intentado modelar el precio de casas de distintas localidades. Entre ellas, la investigación de Lu & et al. (2017) en donde usaron máquinas de vectores de soporte, XGBoost, Regresión Lasso y Ridge, además de regresiones híbridas para predecir el precio de casas. Aquí, para tener el supuesto de normalidad necesario para algunos modelos, aplican una transformación logarítmica, lo cual podría hacer perder un poco de interpretabilidad a los resultados finales. Luego, el mejor modelo fue un modelo híbrido con 65 % de regresión Lasso y 35 % de XGBoost.

En consiguiente, se estudia la investigación de Muralidharan & et al. (2018) en la cual se usan redes neuronales, árbol de decisión y regresión lineal múltiple para predecir el precio de casas en Boston, Estados Unidos. En esta investigación se destaca la diversidad de algoritmos que se pueden realizar para predecir precios de casas, en donde se debe mediante métricas de desempeño poder evaluar estos distintos modelos. Sin embargo, hubo poco análisis y tratamiento de los datos en este estudio, lo cual pudo haber afectado a los resultados de este. Además, al aplicar ciertos modelos, como el de regresión lineal múltiple, se realizan ciertos supuestos como es el de normalidad en los residuos, el cual no es siempre apto para este tipo de datos, lo cual se busca evitar en el presente trabajo.

Finalizando, se estudia el estudio de Shahhossei & et al. (2019) en el cual utilizan regresión Lasso, Random Forest, Redes neuronales, XGBoost y máquinas de vectores de soporte con distintos kernels, además de un diseño de optimización de pesos de todos los modelos propuestos para predecir el precio de casas en Boston y Ames, Estados Unidos. Aquí, el mejor modelo por sí solo para el precio de casas en Boston fueron los de machine learning XGBoost y Random Forest, mientras que para Ames fue la regresión Lasso y Random Forest. Sin embargo, el diseño de optimización incluyendo todos los modelos fue el que mejor resultados tuvo en las 2 localidades, dejando en claro una nueva variedad de métodos que se podrían usar para ajustar este fenómeno.

Dado los estudios vistos, se cree que es necesario implementar un nuevo método, específicamente el de regresión lineal generalizado con respuesta Gamma, debido a los supuestos cuestionables que realizaron las investigaciones previas. Para poder aplicar esto, Hardin & Hilbe. (2018) señalan que los modelos lineales generalizados Gamma puede ser usado cuando la variable de respuesta es continua y puede tomar valores mayores a 0. Además, funciona mejor cuando hay un coeficiente de variación constante. Además, una de las ventajas de este modelo que es bastante flexible. Dado lo expuesto por Hardin & Hilbe, se busca comprobar estos requisitos, para poder tener un respaldo importante a la hora de aplicar el modelo propuesto.

Capítulo 3

Marco teórico

3.1. Modelo lineal generalizado con respuesta Gamma

A continuación se define la componente aleatoria, componente sistemática y la función de enlace de un modelo lineal generalizado con respuesta Gamma (McCullagh y Nelder, 1989):
La componente aleatoria se define suponiendo que $Y_1, Y_2, ..., Y_n$ son variables aleatorias i.i.d. tales que:

$$Y_i \sim \Gamma(\mu_i; \phi), \quad i = 1, 2, ..., n. \tag{3.1}$$

Es decir, se asume que estas variables no necesariamente tienen la misma media y tienen coeficiente de variación constante, $\phi^{-1/2}$.

En cuanto a la componente sistemática, se presenta a continuación:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} = \boldsymbol{X_i^T \beta} \tag{3.2}$$

Donde $\boldsymbol{X_i^T}$ contiene los valores de las variables explicativas y $\boldsymbol{\beta}$ es un vector de parámetros de interés.

Finalizando, la función de enlace relaciona la componente aleatoria y sistemática mediante la siguiente función:

$$g(\mu_i) = \eta_i \tag{3.3}$$

La función de enlace canónica se define como:

$$\frac{1}{\mu_i} = \eta_i \tag{3.4}$$

3.2. Devianza

En el proceso de ajustar un modelo ocurren discrepancias entre los valores ajustados y los valores observados. Medir esta discrepancias para modelos lineales generalizados se trabaja primordialmente con lo formado a partir del logaritmo de una razón de verosimilitud, la que se denomina devianza. Así, los distintos modelos se pueden comparar utilizando estadísticas basadas en la devianza, donde el que tenga menor valor es el mejor modelo (Hardin & Hilbe, 2018).

Específicamente, la devianza para el caso Gamma se expresa de la siguiente forma (McCullagh y Nelder, 1989):

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left[-log\left(\frac{\hat{\mu_i}}{y_i}\right) + \frac{y_i - \hat{\mu_i}}{\hat{\mu_i}}\right] \tag{3.5}$$

Con $\hat{\mu_i} = g^{-1}(\hat{\eta_i})$ y $\hat{\eta_i} = \boldsymbol{X_i^T}\hat{\boldsymbol{\beta}}$.

3.3. Criterio de Información de Akaike

El criterio de información de Akaike, AIC por sus siglas en inglés, fue presentado por Akaike (1974). Este criterio se basa en la estimación relativa de la información perdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos. Este resultado no da información acerca de la calidad del modelo en sentido absoluto, es decir, si todos los modelos candidatos encajan mal, AIC no dará ningún aviso de ello.

Así, el AIC en el caso general está dado por:

$$AIC = 2p - 2ln(L) \tag{3.6}$$

donde p es el número de parámetros en el modelo estadístico, y L es el máximo valor de la función de verosimilitud para el modelo estimado. Entonces, AIC penaliza con una función creciente la cantidad de parámetros que tiene el modelo en cuestión.

Entonces, dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC.

3.4. Residuos

Según Hardin & Hilbe (2018) los residuos del modelo son simplemente la diferencia entre los valores reales y predichos por el modelo, es decir:

$$r_i^R = y_i - \widehat{\mu_i} \tag{3.7}$$

Luego, Hardin & Hilbe (2018) señalan que debe cumplirse 3 supuestos importantes de los residuos, los cuales son:

1. Independencia
2. Varianza constante
3. Media igual a 0

También señalan que existen varios tipos de residuos que se pueden calcular a partir de los residuos originales. En este estudio, se tomarán 2 tipo de residuos y se analizarán las respectivas gráficas de estos:

- Residuos obtenidos
- Residuos de Pearson

En cuanto a los residuos de Pearson, Hardin & Hilbe (2018) definen que este tipo de residuos sirven para poder analizar de mejor manera la independencia y varianza, pues se realiza un escalamiento de los residuos para dejarlos en una escala entre -1 y 1 aproximadamente. Su calculo, es el siguiente:

$$r_i^P = \frac{y_i - \widehat{\mu_i}}{\sqrt{v(\widehat{\mu_i})}} \tag{3.8}$$

Notando que simplemente se divide por la desviación estándar de los residuos.

Dado lo anterior, en este trabajo se revisarán ambos gráficos residuales para poder obtener información acerca de los supuestos residuales que debe cumplir el modelo para ser óptimo.

Capítulo 4

Metodología

4.1. Base de datos

El conjunto de datos a trabajar contiene 1460 observaciones y 13 variables, estas son diferentes características y calificaciones que a priori influyen en el valor de una propiedad. En este caso, las descripciones de casas a analizar serán de la ciudad de Ames, Estados Unidos. Las variables que contiene la base de datos son:

- OverallQual: Califica el material general y acabado de la casa (1-10).
- OverallCond: Califica el estado general de la casa (1-10).
- BsmtFullBath: Baños completos en sótano.
- FullBath: Baños completos sobre rasante.
- HalfBath: Medios baños sobre rasante.
- TotRmsAbvGrd: Habitaciones totales sobre rasante.
- Fireplaces: Número de chimeneas.
- LotFrontage: Pies lineales de calle conectados a la propiedad.
- LotArea: Pies cuadrados de la casa.
- GrLivArea: Pies cuadrados de superficie habitable sobre el nivel del suelo.
- YearBuilt: Fecha original de construcción.
- YearRemodAdd: Fecha de remodelación.
- SalePrice: Precio de la casa.

En cuanto a las variables marcadas con color rojo, LotFrontage, LotArea y GrLivArea se cambió la escala de medida de pies a metros, debido a que es una métrica que se identifica mejor y así poder tener mejores interpretaciones. Para lograr lo anterior, fue necesario multiplicar por $0,3048$ los valores para la variable de metros lineales y por $0,3048^2$ para las variables de metros cuadrados, logrando convertir la medida de pies a metros y pies cuadrados a metros, respectivamente.

Por otro lado, las variables YearBuilt y YearRemodAdd se cambiaron la escala de fechas de creación y remodelación por años de antigüedad de estos acontecimientos, para también interpretar mejor los resultados. Para aquello, fue necesario restar 2022 (año de la presente investigación) con la fecha de antigüedad y de remodelación.

Dado lo anterior, en lo que resta de informe se comentarán estas variables según los cambios realizados. Luego, es necesario poder entender de mejor manera el comportamiento de cada variable presentada, es por eso que a continuación se presenta una tabla descriptiva de las variables enunciadas, considerando el mínimo, mediana, máximo, media y desviación estándar:

Variable	**Mínimo**	**Mediana**	**Máximo**	**Media**	**desv. est.**
OverallQual	1	6	10	6.09	1.38
OverallCond	1	5	9	5,57	1,11
BsmtFullBath	0	0	3	0,42	0,51
FullBath	0	2	3	1,56	0,55
HalfBath	0	0	2	0,38	0,50
TotRmsAbvGrd	2	6	14	6,51	1,62
Fireplaces	0	1	3	0,61	0,64
LotFrontage	6,40	21,03	95,40	21,35	7,40
LotArea	120,8	880,6	19.996,9	977	927,28
YearBuilt	12	49	150	50,73	30,20
YearRemodAdd	12	28	72	37,13	20,64
GrLivArea	31,03	136,01	524,16	140,79	48,81
SalePrice	34900	163000	582933	180502	77086,49

Cuadro 4.1: Estadísticas descriptivas de variables consideradas en la investigación. Elaboración propia.

Se puede observar que según la mediana, al menos el 50 % de las casas que se van a analizar tienen una calificación de 6 en una escala del 1 al 10 en el material general y acabado de la vivienda con un mínimo y máximo de 1 y 10, respectivamente, además, en promedio también se obtiene una calificación de 6, con una desviación estándar de 1,38. Lo anterior, indica que en general las

casas tienen una calificación medio alto, lo cual debería influir en el precio de estas. En el caso de la calificación del estado general de las casas disminuye la mediana a 5, incluso ninguna casa de la base de datos obtiene una calificación del máximo 10. También se disminuye la media y la desviación estándar, por lo cual se concluye que en este apartado, en relación con la calificación del material general y acabado de la casa, disminuye en general la calificación pero es menos variable, posiblemente debido a que se toman más características para evaluar esta nota.

Por otro lado, se observa que al menos el 50 % de las casas no tiene baños en el sótano, lo cual podría considerarse normal, destacando que hay una porción de casas que tiene hasta 3 baños en el sótano. Se destaca que la media es 0,42, la cual se aproximaría a 0 baños en el sótano y desviación estándar de 0,51, lo cual indica que en general las casas no tendrían baños en el sótano. En cuanto a los baños completos de la casa se aprecia, sorprendentemente, que hay una cantidad de casas con 0 baños, lo cual es bastante raro. Según la mediana, hay al menos un 50 % de casas que tienen 2 baños, lo cual es lógico según las características de la ciudad de Ames, con un máximo de 3. La media y desviación estándar indican que en general las casas tienen entre 1 y 2 baños. Por último, en cuanto a los medios baños en la casa, al menos el 50 % de las casas no tiene este tipo de baños, con una media de 0,38 y desviación estándar de 0,5, por ende la proporción de casas con este tipo de baños no es alta.

En cuanto a las habitaciones totales sobre rasante, se aprecia que hay una importante diferencia entre el mínimo y máximo, de 2 hasta 14 habitaciones en algunas casas, lo cual es lógico pues, al ser una zona de muchas universidades, existen varias pensiones especiales para estudiantes. Se aprecia también que al menos un 50 % de las casas tiene 6 habitaciones, nuevamente esto puede ser lógico por las características de la zona.

Por otro lado, según las estadísticas del número de chimeneas, se aprecian resultados normales, con un mínimo de 0 y máximo de 3 chimeneas y con al menos un 50 % de casas con 1 chimenea.

Luego, en los metros lineales de calle conectados a las casas, se observa una importante diferencia, debido a que existe una casa que tiene 6,4 metros lineales de calle en comparación con la que tiene más que son 95,4. Esta diferencia puede ser sustancial al momento de definir los precios de una casa. También se observa una mediana y media de 21 metros lineales lo cual es bastante normal, con una desviación estándar de 7,4, indicando un grado notorio de variabilidad. En esta misma línea, en los metros cuadrados de la casa, que también es una variable importante en la influencia del valor de una propiedad, se observa que hay al menos una propiedad de 19.996 metros cuadrados, una vivienda con bastante espacio en comparación con la que tiene menos que es de 120 metros cuadrados. Se aprecia también que al menos el 50 % de las casas tienen 880 metros cuadrados, lo cual indica que hay muchas casas con gran espacio, también se destaca

la media que es de 977 y desviación estándar de 927,28 lo cual indica un grado de variabilidad bastante notorio, que seguramente es causado por las casas que tienen metros cuadrados extremos. En consiguiente, según los metros cuadrados de superficie habitable en el sueño, se observa que baja bastante (es lógico) siendo un mínimo de 31 metros cuadrados y máximo de 524, con una mediana de 136,01. Estas variables pueden ser muy importantes para determinar el precio de casas, pues entre más grande sea una casa, mayor debería ser el precio de esta.

En cuanto a las variables de año de construcción y de remodelación de la casa, se aprecia que las dos tienen el mismo mínimo y es lógico, pues hay casas que no tienen remodelaciones y, por tanto, en este caso el año de antigüedad de construcción y remodelación es el mismo. Se aprecia que el 50 % de las casas tiene una antigüedad de 49 años con un máximo de 150, lo cual habla bastante de que esta ciudad tiene una proporción de casas creadas antes del siglo 21. También se observa que según los años de remodelación, al menos el 50 % de las casas remodeló la casa hace 28 años, lo cual también indica que la mayoría tiene su última remodelación antes del siglo 21. Estas variables pueden ser muy importantes también, pues a medida que la casa es más antigua, debería bajar el precio de esta, a menos que sea una casa con una tradición o historia particular.

Finalizando, la última variable y la más significativa en el trabajo es la del precio de la casa, la cual es la que se busca modelar a través de las demás variables, la cual presenta un mínimo de precio de 34.900 dólares, lo cual se puede considerar accesible para una persona de clase media. En cuanto al máximo, se observa que la casa más cara cuesta 582.933 dólares, lo cual es bastante. Además, tiene una mediana de 163.000 dólares, lo cual indica que al menos el 50 % de las casas está sobre este precio, siendo un valor alto para una familia de clase media (al menos en Chile).

4.1.1. Visualización de variables cualitativas

En consiguiente, se presentan distintos gráficos para analizar de mejor manera la distribución de las distintas variables. Comenzando, se analizan primero las variables cualitativas, empezando por la calificación de material y acabado de la casa, calificación de estado general de la caja, baños completos en sótano y baños completos de la casa como se ve en la figura 4.1: Es notorio que en

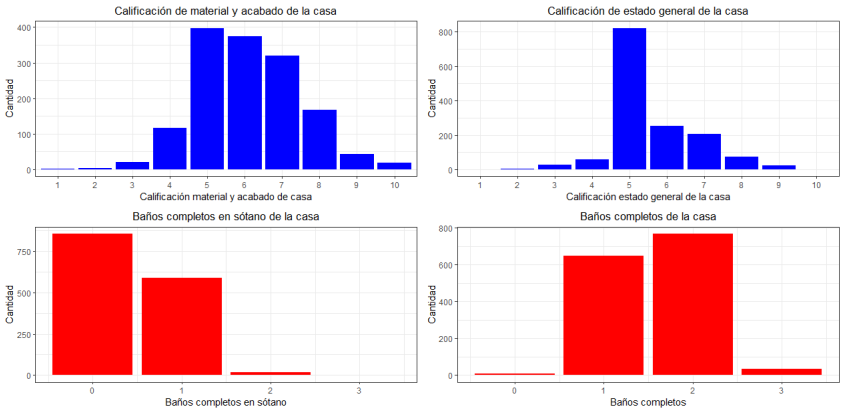

Figura 4.1: Gráficos de barra de la calificación de material y acabado de la casa y estado general de la casa, baños completos en sótano y completos en total. Elaboración propia.

las calificaciones de material y acabado de la casa existe al comienzo un aumento de frecuencias a medida que aumenta la nota, para luego desde la nota 6 van bajando las cantidades. Se concentran los datos mayoritariamente entre las notas 5, 6 y 7. En la calificación del estado general de la casa se aprecia que las notas se concentran mayoritariamente en la nota 5, para luego distribuir de manera casi equitativa las frecuencias en el resto de notas. En cuanto a los baños completos en sótano, se aprecia que la gran mayoría de casas tiene 0 o 1 baños en sótano, dejando una cantidad mucho menor de casas con 2 o 3 baños en sótano. En cuanto a los baños completos de la casa, se aprecia que la gran mayoría tiene 1 o 2 baños completos en casa lo cual es bastante lógico, por otro lado, hay una muy poca cantidad de casas con 0 baños en casa, lo cual aunque pareciese raro, puede ocurrir.

Ahora, se presentan unos gráficos de barra sobre las variables medios baños en la casa, chimeneas en la casa y habitaciones en la casa:

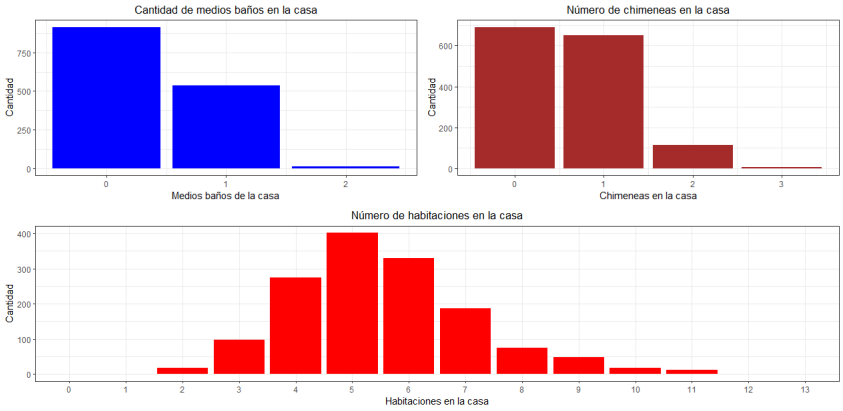

Figura 4.2: Gráficos de barra de canntidad de medios baños, número de chimeneas y habitaciones en la casa. Elaboración propia.

Se observa que la gran mayoría tiene 0 medios baños en la casa, lo cual es bastante lógico, y una frecuencia de aproximadamente 500 tiene 1 medio baño. El número de chimeneas en la casa, tiene una proporción parecida cuando tienen 0 y 1 chimeneas, para luego bajar esta frecuencia bastante para 2 o 3 chimeneas por casa. Finalizando, las habitaciones en las casas varían mayoritariamente entre 4 y 7 habitaciones, lo cual puede ser producto por las características de la ciudad.

4.1.2. Visualización de variables cuantitativas

Continuando con la visualización, ahora se presentan algunos gráficos de las variables cuantitativas, empezando por los metros lineales de calle conectados a la casa:

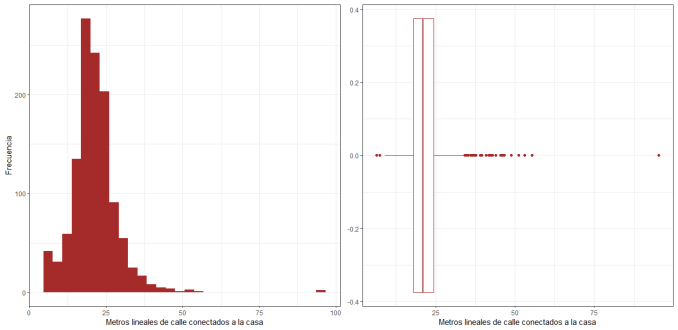

Figura 4.3: Histograma y Box-Plot de la variable de metros lineales de calle conectados con la casa. Elaboración propia.

Se aprecia que su distribución es parecida a la de una distribución Chi cuadrado o Gamma, con colas bastante pesadas y con valores atípicos, además, según el Box-Plot, se observa una simetría en la distribución de los datos sin considerar los valores atípicos. Estos valores atípicos no se decide tratarlos, debido a que aportan información valiosa para los modelos, pues estos pueden ser factores importantes al momento de determinar el precio de una casa.

Luego, se presenta un histograma y Box-Plot de los metros cuadrados de la casa:

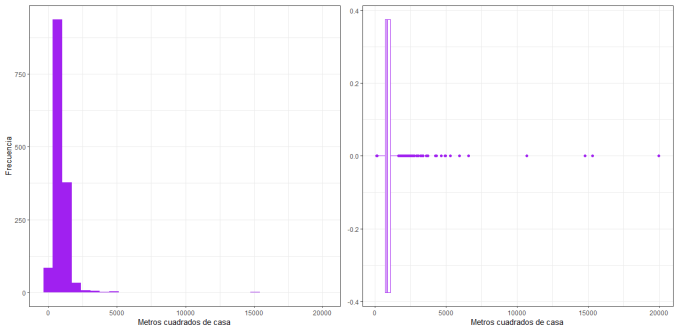

Figura 4.4: Histograma y Box-Plot de la variable de metros cuadrados de casa. Elaboración propia.

En la figura 4.4 es notorio que tiene valores atípicos muy alejados, además tiene muchos valores concentrados alrededor de los 1000 metros cuadrados de casa. Esta variable es muy importante, por lo que estos valores atípicos tampoco se decide tratarlos, ya que pueden definir notoriamente el precio de alguna casa.

En consiguiente, se presenta el histograma y Box-Plot de los metros cuadrados de superficie habitable de las casas:

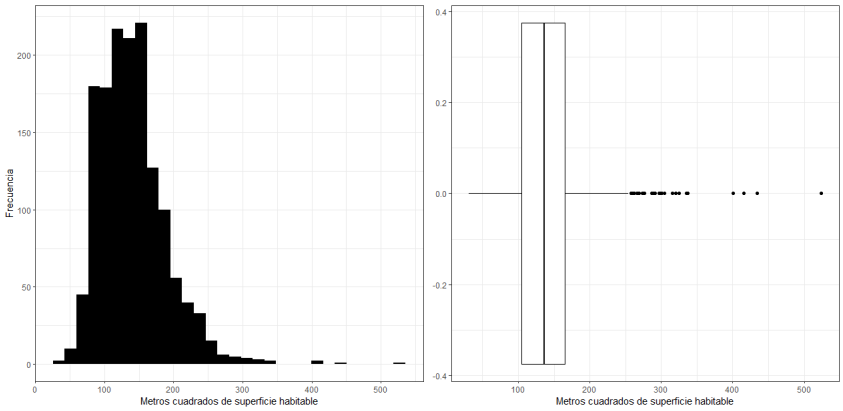

Figura 4.5: Histograma y Box-Plot de metros cuadrados de superficie habitable de casa. Elaboración propia.

Según el histograma, la distribución de esta variable es parecida a la de una distribución Gamma o Chi cuadrado, teniendo una subida notoria y luego una cola pesada a la derecha. Se observan también algunos valores atípicos, característica natural de este tipo de variables. También estos valores atípicos son importantes para el ajuste del precio de las casas, por lo que no se tratan.

Luego, se presentan los histogramas y Box-Plot de los años de antigüedad y años desde la última remodelación de las casas, como se muestra en la figura 4.6:

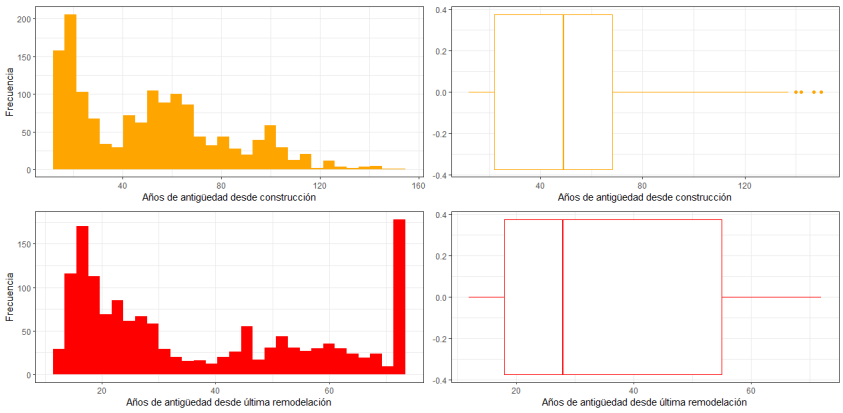

Figura 4.6: Histograma y Box-Plot de los años de antigüedad y años desde la última remodelación de las casas. Elaboración propia.

Se observa que las dos variables tienen una distribución parecida, a excepción de la dispersión de estas variables, pues las de años de antigüedad es más estable que la de años desde la última remodelación, pues esta última tiene valores bastante concentrados sobre los 100, posiblemente porque sean casas antiguas que nunca recibieron alguna remodelación.

4.1.3. Variable de respuesta: Precios de casas

En cuanto a la variable de respuesta, es necesario entender la distribución de esta variable, pues esto es información valiosa para aplicar el modelo lineal generalizado Gamma. Para ello, se presenta a continuación en la figura 4.7 un histograma y Box-Plot de esta variable:

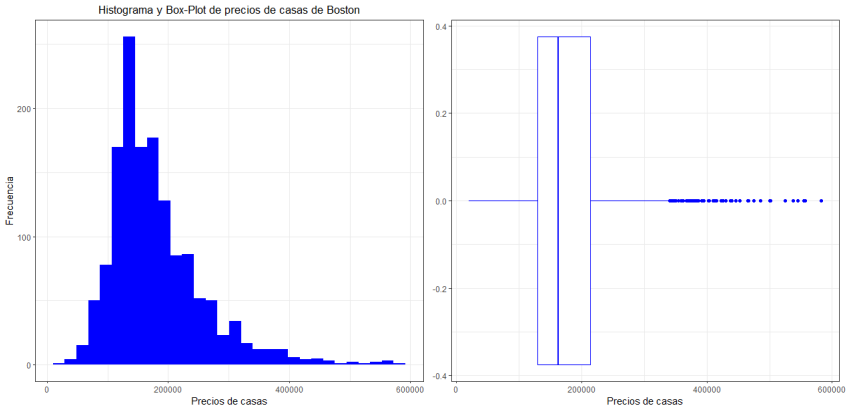

Figura 4.7: Histograma y Box-Plot de los precios de casas. Elaboración propia.

Se observa una distribución bastante parecida a una distribución Gamma, con asimetría positiva y varios valores atípicos que pertenecen a la cola de la derecha, lo cual es bastante común en este tipo de distribución. Visualmente, es bastante bueno para la aplicación del modelo lineal generalizado Gamma.

Luego, para comprobar si tiene coeficiente de variación constante, dado que si es así es un apoyo mayor a aplicar un modelo lineal generalizado Gamma, se presenta en la figura 4.8 un gráfico del histograma de esta variable, pero partido en 3 tramos, primero en la curva ascendente de los datos, luego en la que desciende y por último la de los valores de la cola de la derecha, cada tramo con su respectivo coeficiente de variación:

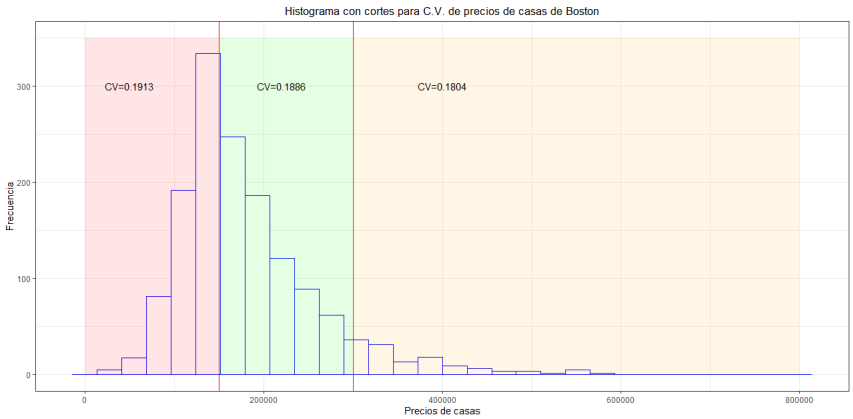

Figura 4.8: Histograma de los precios de casas junto con el coeficiente de variación para algunos tramos. Elaboración propia.

Se aprecia que el coeficiente de variación es bastante cercano en los 3 tramos elegidos, por lo cual se considera que esta métrica es constante en la variable. Este resultado es bastante bueno para los supuestos del modelo lineal generalizado Gamma, pues como se mencionó anteriormente, el modelo funciona mejor si se cumple este requisito.

4.1.4. Correlación entre variables

Es posible que exista multicolinealidad entre las covariables, por lo que se presenta una matriz de correlación para analizar esta incógnita, debido a que esto puede ser negativo para los resultados de los modelos propuestos, pues puede afectar el valor de los coeficientes, la variabilidad de estos, e incluso haber infinitas soluciones para los coeficientes si esta multicolinealidad es perfecta, debido a que una o más variables estarían aportando la misma información al modelo. Esta matriz es la más común para poder estudiar este caso, ya que es bastante explícita y fácil de entender. Aparte de entregar la información directamente. Para calcular la matriz de correlación, se separaron las variables cualitativas y cuantitativas para aplicar distintos métodos según el tipo de variable. Para el caso de las variables cualitativas, se ocupa el método de Spearman, obteniendo el siguiente resultado:

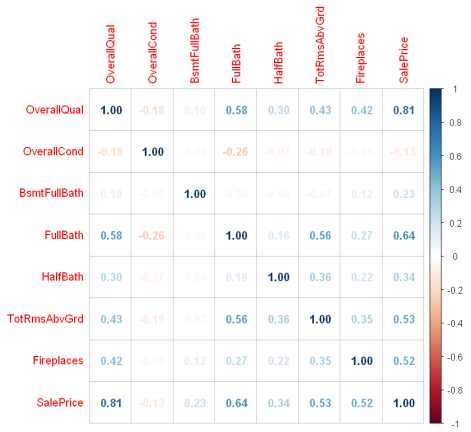

Figura 4.9: Correlación de Spearman para variables cualitativas.

Se aprecia que entre las covariables no existen correlaciones muy altas, siendo la más fuerte con 0,58 la de baños completos de la casa con la calificación del material general y acabado de la casa, por ende no se aprecia que pueda existir problemas de multicolinealidad en estas variables. En comparación con la variable de respuesta de precios de casas, se ve que la mayoría tiene un cierto grado de correlación, sin embargo, la calificación del estado general de la casa con baños completos en sótano no se observan una correlación tan fuerte.

Por otro lado, para calcular la correlación de las variables cuantitativas, se usa el coeficiente de correlación de Pearson, obteniendo los siguientes resultados:

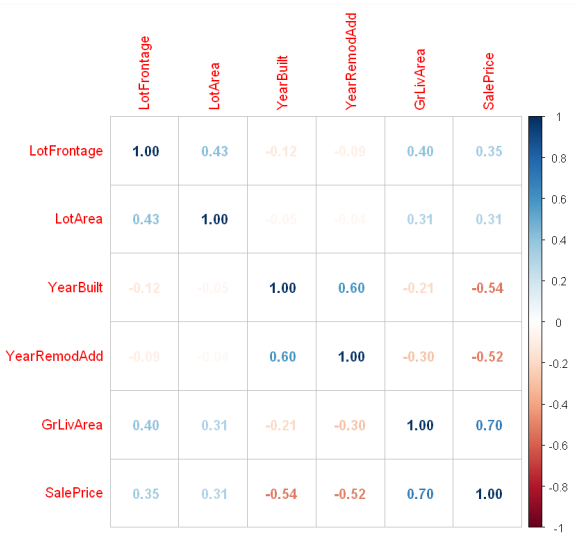

Figura 4.10: Correlación de Pearson para variables cuantitativas.

Al igual que el anterior caso, no se presentan correlaciones tan fuertes, siendo la más fuerte la de años de antigüedad de la casa con los años de antigüedad de remodelación de la casa con 0,6 de correlación, lo cual es bastante lógico, pues estas variables son indicadoras de características parecidas. En cuanto a la variable de respuesta, todas las covariables presentan algún grado de correlación con esta, por lo que es bastante bueno para los modelos.

4.2. Aplicación de modelos

Se eligieron los mejores modelos Gamma con sus distintos enlaces, según la función de R *bestglm* proveniente de la librería *bestglm* según AIC y BIC. Luego de obtener los mejores modelos, se hace el ajuste de estos, en donde el resultado de los coeficientes del modelo Gamma con enlace identidad se presenta a continuación:

β_i	$\widehat{\beta}_i$	**error estándar**	**Estadístico T**	**Valor p**
β_0	21162.139	5199.062	4.070	$< 0,01$
LotFrontage	637.475	147.486	4.322	$< 0,01$
LotArea	18.614	2.497	7.454	$< 0,01$
OverallQual	11971.167	772.825	15.490	$< 0,01$
OverallCond	3226.253	672.534	4.797	$< 0,01$
YearBuilt	-519.825	34.554	-15.044	$< 0,01$
YearRemodAdd	-256.475	43.714	-5.867	$< 0,01$
GrLivArea	606.616	39.242	15.458	$< 0,01$
BsmtFullBath	10060.494	1573.969	6.392	$< 0,01$
TotRmsAbvGrd	-3727.073	874.287	-4.263	$< 0,01$
Fireplaces	10613.966	1570.336	6.759	$< 0,01$

Cuadro 4.2: Resultados de los coeficientes estimados $\widehat{\beta}_i$ mediante el modelo Gamma con enlace identidad. Elaboración propia.

Se aprecia en el cuadro 4.2 que el mejor modelo no consideró las variables de medios baños sobre rasante (*HallBath*) y baños completos sobre rasante (*FullBath*). Luego, según los coeficientes estimados obtenidos del modelo, es notorio que todos son distantes del 0 por mucho, tan así, que el valor p obtenido en todos los casos fue menor a 0,01. Lo anterior, indica que según la evidencia obtenida, hay una probabilidad muy baja de cometer el error tipo 1 y, por tanto, se rechaza la hipótesis nula de que los coeficientes sean iguales a 0, con un nivel de confianza mayor al 99 %. En cuanto al error estándar de los coeficientes, se observa en algunos un grado de variabilidad alta.

Ahora, se presenta la misma tabla pero para el modelo Gamma con enlace logarítmico:

β_i	$\widehat{\beta_i}$	**error estándar**	**Estadístico T**	**Valor p**
β_0	10.766240177	0.042212680	255.048	$< 0,01$
LotFrontage	0.003502163	0.000673338	5.201	$< 0,01$
LotArea	0.000011799	0.000002017	5.850	$< 0,01$
OverallQual	0.105935141	0.005004673	21.167	$< 0,01$
OverallCond	0.049183831	0.004746920	10.361	$< 0,01$
YearBuilt	-0.003732138	0.000222981	-16.737	$< 0,01$
YearRemodAdd	-0.001370450	0.000291732	-4.698	$< 0,01$
GrLivArea	0.000876570	0.000038232	22.927	$< 0,01$
BsmtFullBath	0.081196315	0.008529147	9.520	$< 0,01$
Fireplaces	0.058328048	0.008050424	7.245	$< 0,01$

Cuadro 4.3: Resultados de los coeficientes estimados $\widehat{\beta_i}$ mediante el modelo Gamma con enlace logarítmico. Elaboración propia.

Se aprecia en el cuadro 4.3 que en este modelo, quedaron fuera las variables *HalfBath* y *FullBath*, nuevamente. En cuanto a los coeficientes estimados, ahora solo el coeficiente β_0 tuvo una estimación lejana notoriamente del 0, el resto todos cercanos a 0. Sin embargo, según el valor p, nuevamente existe evidencia suficiente para rechazar la hipótesis nula de que estos coeficientes son iguales a 0, seguramente influido por el tamaño muestral que se tiene en la base de datos. Además, el error estándar se observa bastante más bajo que en el anterior modelo, lo cual es bastante bueno.

Finalizando, se presenta la misma tabla pero del modelo Gamma con enlace inverso:

β_i	$\widehat{\beta_i}$	**error estándar**	**Estadístico T**	**Valor p**
β_0	0.00001141965058	0.00000032705275	34.917	$< 0,01$
LotFrontage	0.00000000898526	0.00000000272557	3.297	$< 0,01$
LotArea	-0.00000000004030	0.00000000000839	-4.803	$< 0,01$
OverallQual	-0.00000079529870	0.00000002865454	-27.755	$< 0,01$
OverallCond	-0.00000026353097	0.00000004272249	-6.168	$< 0,01$
YearBuilt	0.00000001748472	0.00000000196422	8.902	$< 0,01$
YearRemodAdd	0.00000001597085	0.00000000271155	5.890	$< 0,01$
BsmtFullBath	-0.00000018173190	0.00000005944000	-3,057	$< 0,01$
Fireplaces	-0.00000055328181	0.00000005328097	-10.384	$< 0,01$

Cuadro 4.4: Resultados de los coeficientes estimados $\widehat{\beta_i}$ mediante el modelo Gamma con enlace inverso. Elaboración propia.

En el cuadro 4.4 se observan resultados parecidos al modelo Gamma con en-

lace logarítmico, pero ahora el coeficiente estimado β_0 también ahora es cercano a 0. Nuevamente, se tienen valores p menores a 0,01 y errores estándar parecidos.

Luego de aplicar los modelos, se grafica la densidad de los valores que predice con los valores reales. Se grafica la densidad de estos debido a que permite suavizar el denominado "ruidoz así poder observar de mejor manera el comportamiento de ambas. Se presenta el gráfico de lo obtenido mediante el modelo Gamma con enlace identidad:

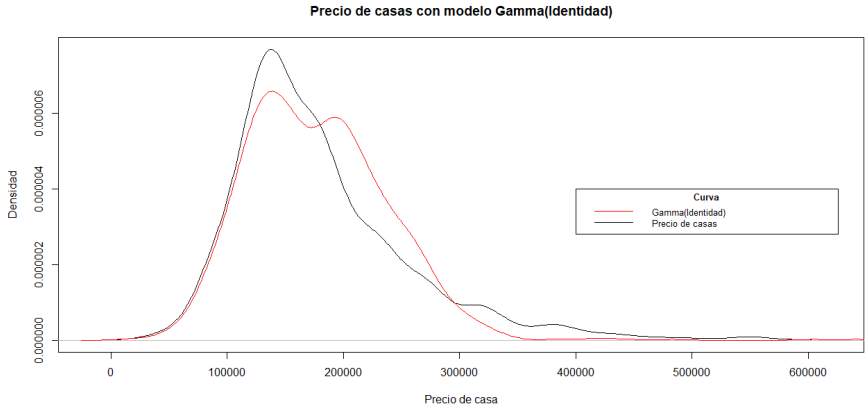

Figura 4.11: Densidad de modelo Gamma con enlace identidad, junto con la densidad de los valores reales de precios de casas. Elaboración propia.

En la figura 4.11 se observa que si bien al principio se ajusta bastante bien el modelo, cuando llega al precio de 120000 aproximadamente, empieza el ajuste a alejarse de los valores reales, no logrando llegar al peak de la densidad de los precios reales que ocurre cuando el precio es 150000 apróx, subestimándolo, luego, cuando llega al precio de 200000 no sigue la bajada que sufre la distribución, sino que aumenta un poco para luego si bajar. Ya llegando a la cola de la distribución, subestima los precios en el rango entre 300000 y 400000 notoriamente, para luego asemejarse de mejor manera a la cola.

Ahora, se presenta el mismo gráfico pero ahora del modelo Gamma con enlace logarítmico:

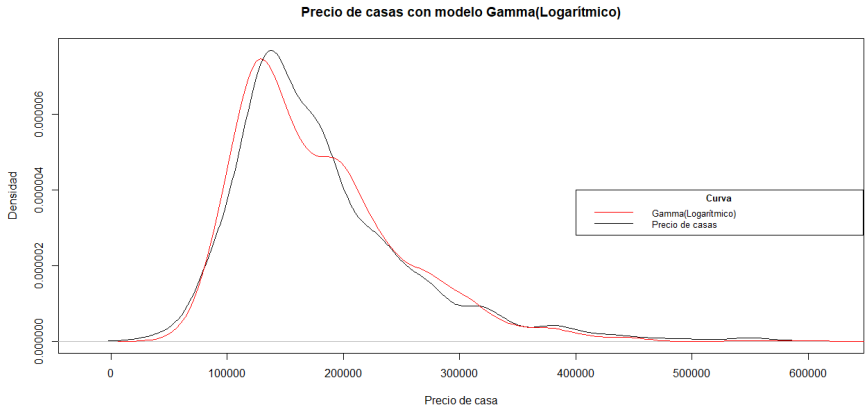

Figura 4.12: Densidad de modelo Gamma con enlace logarítmico, junto con la densidad de los valores reales de precios de casas. Elaboración propia.

En la figura 4.12 se aprecia un mucho mejor ajuste del modelo, logrando casi llegar al peak de densidad en los precios de 150000, sin embargo, luego del peak subestima los precios notoriamente hasta el precio de 200000. En general, se ajusta bastante bien el modelo, a excepción de este trazo.

Finalizando, se presenta el mismo gráfico pero ahora con el modelo Gamma con enlace inverso:

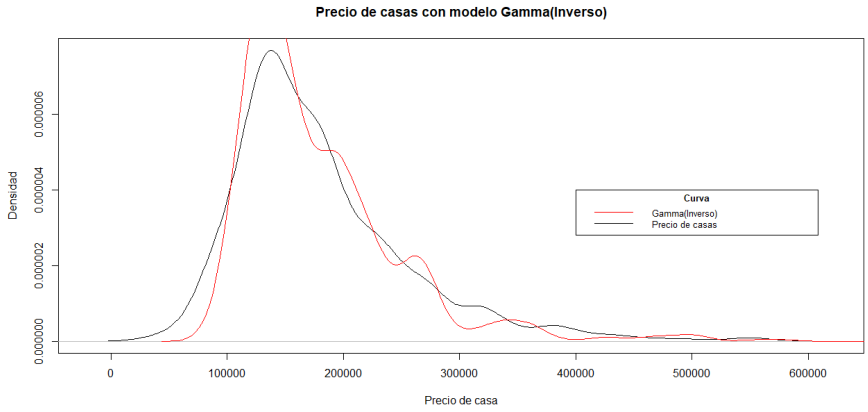

Figura 4.13: Densidad de modelo Gamma con enlace inverso, junto con la densidad de los valores reales de precios de casas. Elaboración propia.

En la figura 4.13 se aprecia un ajuste mucho más volátil que los 2 anteriores, además, el modelo sobreestima en demasía el tramo de precios entre 100000 y 170000. Luego, tiene fluctuaciones poco lógicas entre los precios de 180000 y 300000.

En conclusión, según los gráficos vistos, el modelo Gamma con enlace logarítmico pareciera ser el que mejor ajuste obtuvo, mientras que el modelo Gamma con enlace inverso es el de peor resultados.

4.3. Análisis de bondad de ajuste

Para analizar la bondad de ajuste de los distintos modelos, se calcula la devianza y AIC de cada uno que se resume en la siguiente tabla:

Modelo	**Devianza**	**AIC**
Gamma(Identidad)	210.522	28088
Gamma(Logarítmico)	210.520	27873
Gamma(Inverso)	210.522	28509

Cuadro 4.5: Resultados de Devianza y AIC para los 3 modelos propuestos. Elaboración propia.

Se observa que el modelo Gamma con enlace logarítmico obtuvo una menor devianza, aunque solo por 2 milésimas, lo cual no es una diferencia notoria, también obtuvo un menor valor de AIC. Según estos resultados, el modelo Gamma con enlace logarítmico es el que mejor bondad de ajuste tiene.

Se destaca, también, que el modelo Gamma con enlace identidad y también el con enlace inverso obtuvieron el mismo valor de devianza, pero en cuanto al AIC el modelo con enlace identidad obtuvo un menor resultado y, por tanto, es mejor que el inverso en esta métrica.

4.4. Análisis residual

Dado el análisis de bondad de ajuste, es necesario revisar si los residuos cumplen los supuestos de media igual a 0, independientes y varianza constante. Para ello, primero se presenta el gráfico de residuos de los 3 modelos. Primero, para el compuesto de la familia Gamma con enlace identidad:

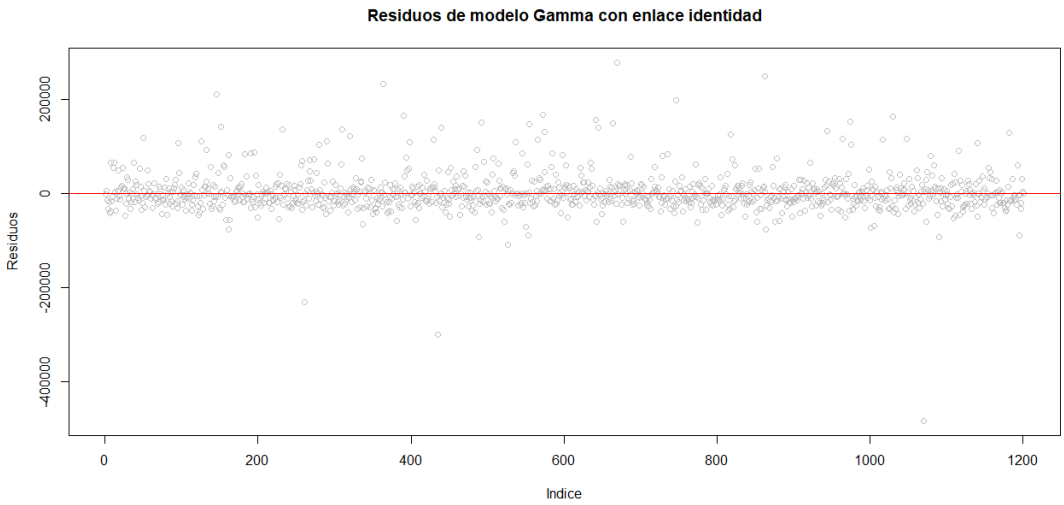

Figura 4.14: Gráfico residual para modelo Gamma con enlace identidad. Elaboración propia.

Se observa según la figura 4.14 que el modelo Gamma con enlace identidad en general los residuos son independientes, pero no se observa una media igual a cero ni tampoco una varianza constante, pues tienen muchos residuos distantes del resto, obteniendo valores incluso sobre -200000 y 200000. Los resultados residuales son bastante malos para este modelo.

Ahora, se presenta el mismo gráfico para el modelo Gamma con enlace logarítmico:

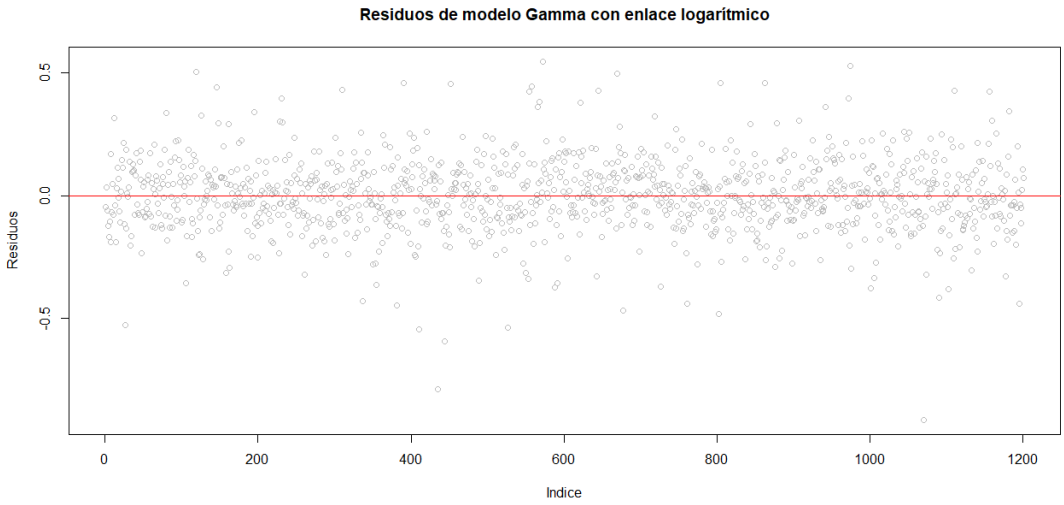

Figura 4.15: Gráfico residual para modelo Gamma con enlace logarítmico. Elaboración propia.

En la figura 4.15 se aprecia que la media de los residuos podría ser 0, hay independencia y la varianza podría ser constante, sin embargo, existen algunos valores que se escapan un poco de los demás, por lo que podría afectar los resultados.

Finalizando, se presenta el gráfico residual del modelo Gamma con enlace inverso:

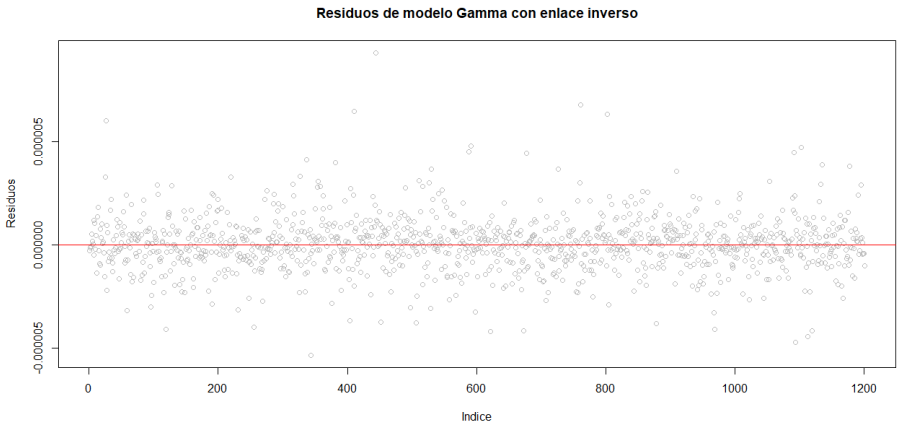

Figura 4.16: Gráfico residual para modelo Gamma con enlace inverso. Elaboración propia.

Se aprecia que tiene algunos valores distantes del resto, pero en general los residuos parecieran tener media 0 e independientes. En cuanto a la varianza, pareciera que también fuera constante, pero los valores distantes podrían afectarla.

Ahora, para poder evitar el análisis sesgado por la escala de los residuos, se analizan los residuos de Pearson para los 3 modelos, presentado en el siguiente gráfico:

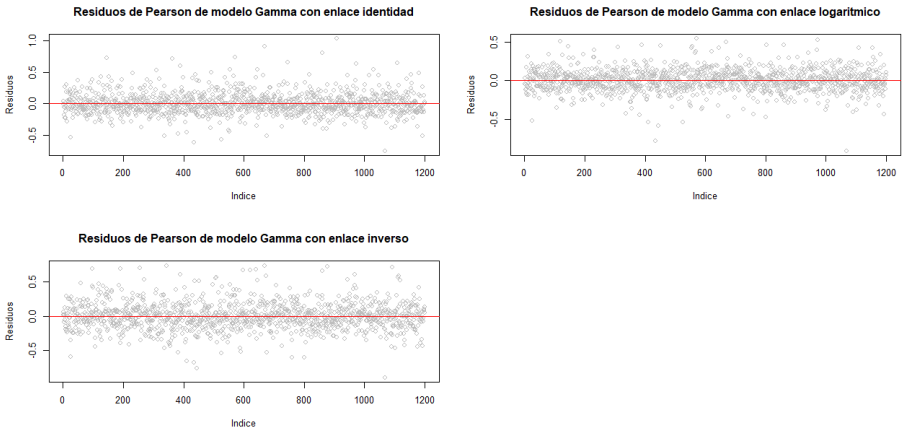

Figura 4.17: Gráfico residual de Pearson para los 3 modelos. Elaboración propia.

Se aprecia que aunque se realizó el escalamiento, igualmente se observa un grado

de variabilidad no constante en los 3 modelos, aunque ahora si se aprecia un poco de independencia.

Según los resultados obtenidos mediante el gráfico de residuos, se puede apreciar que los modelos no cumplen con certeza los supuestos residuales, lo cual es bastante malo, si se puede inferir que el modelo Gamma con enlace logarítmico es el que mejor resultados tuvo.

Capítulo 5

Conclusiones

Según la naturalidad de los datos, es posible que ajustar un modelo lineal Gamma sea lo adecuado, sin embargo, también puede suceder que otros modelos con supuestos un poco distintos, como el modelo Log-normal, también puedan servir para modelar este tipo de problemas. También, según lo visto por resultados anteriores, se pueden obtener buenos resultados aplicando modelos no solo de regresión, sino que también de Machine Learning.

En consiguiente, también hay que considerar que para modelar el precio de casas, en el problema planteado, ha podido faltar información importante para poder realizar esta disyuntiva, como por ejemplo la cantidad de colegios, supermercados, policías, entre otros, cerca de la casa en cuestión, lo cual indudablemente puede ser una variable importante al definir precios de casas. Dado esto, los resultados obtenidos podrían ser mejorados en estudios posteriores, sabiendo también las limitaciones del caso.

En todo caso, dado los factores planteados, ha sido posible ajustar distintos modelos Gamma obteniendo resultados distintos, en donde se destaca que en las pruebas de bondad de ajuste los resultados fueron parecidos, tanto en devianza como AIC y en el análisis residual un poco distantes, pero en donde los 3 modelos propuestos no presentan unos residuos catalogadamente buenos.

Dado lo anterior, si se tuviera que elegir un modelo propuesto, el modelo lineal Gamma con enlace logarítmico ha sido el modelo con mejores resultados, tanto por devianza y AIC, como también por el análisis residual. Este resultado puede ser producto de la suavidad que se obtiene al aplicar logaritmos a alguna variable, debido a que esto hace que la varianza baje.

Capítulo 6

Referencias

- Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE transactions on automatic control, 19(6), 716-723.* Recuperado de: [Ver aquí]

- Hardin, J. & Hilbe, J. (2018). Generalized Linear Models and Extensions. *4ed: Stata Press.* [Ver aquí]

- Lu, S. & et al. (2017). A hybrid regression technique for house prices prediction. *2017 IEEE International Conference on Industrial Engineering and Engineering Management, pp. 319-323.* [Ver aquí]

- Muralidharan, S. & et al. (2018). Analysis and prediction of real estate prices: A case of the Boston housing market. *Issues in Information Systems, 19(2), pp. 109-118.* [Ver aquí]

- McCullagh, P. & Nelder, J. (1989) Generalized Linear Models. *2ed: Chapman and Hall.* [Ver aquí]

- Shahhosseini, M. & et al. (2019). Optimizing ensemble weights for machine learning models: A case study for housing price prediction. *Industrial and Manufacturing Systems Engineering.* [Ver aquí]